# Survey of hypothesis testing on multivariate data

## Multivariate distribution

Let *X* = (*x*1*, x*2*,…, xp*)*T* be *p-*dimension *random vector* with note that *T* denotes transpose operation which changes row to column and otherwise.

Of course X belongs to *p-*dimension space and so vector space mentioned in this report is .

Note that if *X* is *p-*dimension random vector, then its components *xi* (s) are also random variable which is called *partial-variable*. The multivariate distribution function *F* of *X* is defined as the accumulative probability.

*F*(*X*0) *= P*(*X < X*0) = *F*(*x*1 *< x*01*, x*2 *< x*02*,…, xp < x*0*p*)

The density function *f* of *X* is defined as the derivative of accumulative function *F*. In practice, density function is used in lieu of accumulative function.

Probability density function (pdf) satisfies three following conditions.

Note that *R* is the region of random vector *X*, thus, *R* is the *p-*dimension space.

The marginal probability density function of partial-variable *xi* where *xi* is a component of *X* is the integral of *f* over points so that *xi* receives concrete value. Let be the region of *X* so that *xi = a* and so the marginal density function of *xi* is defined as following:

Random partial-variables are mutually independent if and only if the probability density function is the product of marginal density function of partial-variables.

The marginal density function can be extended with *k* partial-variables *xi+1, xi+2,…, xi+k*. Note that it is not necessary for *k* partial-variables to be in successive order such as *i+1, i+2,.., i+k*.

The condition probability of *xj* given a set of *k* partial-variables *xi+1, xi+2,…, xi+k* is the ratio of marginal density function of *xj*, *xi+1, xi+2,…, xi+k* to the marginal density function of *xi+1, xi+2,…, xi+k*.

The partial-variable *xj* is independent from a set of *k* partial-variables *xi+1, xi+2,…, xi+k* if and only marginal density function of *xj*, *xi+1, xi+2,…, xi+k* equals marginal density function of *xj*.

Please pay attention to the concept of independence because it relates to many probabilistic theorems. The condition probability can be extended to *c* partial-variables *xj+1, xj+2,…, xj+c*. Note that it is not necessary for *c* partial-variables to be in successive order such as *j+1, j+2,.., j+c*.

### Some important parameters of multivariate distribution

Now we survey some important parameters of probability distribution such as population mean and population variance. We should distinguish between population parameters and their estimates extracted from sample. For example, sample mean and sample variance are estimates of population mean and population variance, respectively. Population parameters are theoretical parameters which calculated from probability density function. *This section discusses only theoretical parameters*. Given the *p-*dimension random vector *X* = (*x*1*, x*2*,…, xp*)*T*, the partial-mean of partial-variable *xi* is the expected value of *xi*.

Note that theoretical mean is also called expectation or expected value. The conditional mean of variable *xj* given a set of *k* partial-variables *xi+1, xi+2,…, xi+k* is defined as below:

The partial-variance of partial-variable *xi* is:

Note that we prefer notations like , to notations like , . The standard-deviation of *xi* is the squared-root of its variance.

The population partial-covariance between partial-variables *xi* and *xj* is:

We recognize that . The correlation-coefficient of variables *xi* and *xj* is calculated by normalizing their covariance over their standard-deviations. Correlation-coefficient ranges in interval [–1*,* 1]. If it equals 1 then two variables is totally proportional. If it equals –1 then two variables is inversely proportional. If it equals 0 then two variables are uncorrelated.

The mean of random vector *X* is composed of *p* population partial-variances.

The covariance of given random vectors *X* and *Y* is the *p*x*p* matrix whose elements are partial-covariance (s).

We recognize that the covariance ∑*XY* is symmetric matrix. If variables *X* and *Y* are independent, then we infer that *Covar*(*X, Y*) *= 0* but the backward inference is not asserted. The correlation-coefficient matrix of variables *X* and *Y* is composed of partial-correlation-coefficients.

The variance of random vector *X* is defined as the covariance of *X* and itself.

Let *μ = μX* and *∑ = ∑XX*, we denote *X* (*μ,∑*) if *X* has mean *μ* and variance (covariance) *∑*. Let and be constants matrices. Let *X, Y* and *Z* be random vectors. Let *a* and *b* be constant vectors. Let *α* and *β* be scalar numbers. We have some following properties.

### Multinormal distribution

Normal distribution is the heart of hypothesis because almost parametric tests are based on the assumption that sample conforms normal distribution. Multi-dimension normal distribution so-called multinormal distribution is used as an extension of one-dimension normal distribution when sample observations are vectors. Let *μ* and *∑* be mean and covariance matrix, multinormal density function is:

Where |.| denotes determinant of matrix and *exp*(.) denotes exponent function and *∑–1* is the inverse of covariance matrix.

The multinormal density function *N*(*μ,∑*) is constant on ellipsoids following equation (*X – μ*)*T∑–1*(*X – μ*) = *d2*. The half-lengths of axes of contour ellipsoid are where *λi* (s) are eigenvalues of the inverse of covariance matrix *∑–1*.

If random vector *X* conforms multinormal distribution, we denotes *X* *N*(*μ, ∑*). If *μ* =0 is zero vector and is *p*x*p* identity matrix, multinormal distribution becomes standard multinormal distribution or Gaussian distribution.

Note that 0 denotes zero number, zero vector and zero matrix.

According to Jordan decomposition theorem, given symmetric covariance matrix *∑* is decomposed as below:

Where *Γ* is orthogonal matrix and *Λ* is diagonal matrix.

The rational power *∑r/s* of covariance matrix *∑* is defined as below:

Where,

Note that such definition of rational power is applied into any symmetric matrix.

Suppose *X* *N*(*μ,∑*) and let *Z* = , we have *Z* *N*(0*, I*). Otherwise if *X* *N*(0*, I*) and let *Z = ∑*1/2*X + μ*, we have *Z* *N*(*μ,∑*). Suppose *X* *N*(*μ,∑*) and let be non-singular *p*x*p* matrix and let *Y = X + c* be the random vector where *c* is the constants vector, we have:

Suppose *X* *N*(*μ,∑*) and let *U =* (*X – μ*)*T∑*–1(*X – μ*) be random vector, thus, *U* conforms chi-square distribution with *p* degrees of freedom, *U* χ2*p*.

## Multivariate estimation of parameters

Parameters such as mean, variance, median, etc are essential aspects of probability distribution. Previous mentions two important parameters: mean and variance. However parameters are theoretical aspects of distribution and so, they are often estimated from sample. Given a sample, the function of observations is called statistic, for example, sample mean and sample covariance are statistics. Statistical inference includes two methods: parameter estimation and hypothesis testing. Both of them are based on these statistics. Hypothesis testing is main subject in next section and this section focuses on parameter estimation. As aforementioned, parameters are theoretical aspects of distribution and so statistics are used to estimate these parameters, thus, statistics are called estimates of parameters. For example, sample mean , sample covariance are estimates of population mean *μ* and population covariance *∑*, respectively. Note that statistics are considered random variables. When sample is multivariate data, statistics and parameters are vectors or matrix. Hereafter, we browse some concepts of parameter estimation.

### Basic concepts of parameter estimation

Let *θ* and represent multivariate parameter (such as population mean *μ* and population variance *∑*) and it estimate (such as sample mean and sample covariance ). Estimate is unbiased estimate of parameter θ if and only if expectation of equals *θ*.

If is biased estimate, the deviation *E*() – *θ* is called a bias. Let *Var*() be the variance of estimate , the standard-error of denoted *se*() is defined as its standard-deviation.

Given a sample = {*X*1*, X*2*,…, Xn*} consisting of *p*-dimension samples *Xi* (s), the sample mean is the average vector over *n* observation vectors *Xi* (s).

We prove that sample mean is unbiased estimate of theoretical mean *μ*. Suppose *n* observation vectors *Xi* (s) are independent and identically distributed (i.i.d.) with the same theoretical mean *μ*.

The variance of is:

We have:

Similarly, the sample covariance defined as below is unbiased estimate of theoretical covariance *∑*.

Where is the sample partial-covariance between *kth* component and *lth* component of observation vectors. Note that each sample point *Xi =* (*xi1, xi2,…, xip*) is *p*-dimension vector whose components are singular random variable which is also called partial-variable as aforementioned.

Where and are sample partial-means of *kth* component and *lth* component, respectively.

Note that denoted as is sample partial-variance of *jth* component of observation vectors and is the respective sample partial-standard-deviation.

Therefore, sample covariance is re-written as below:

If sample is considered as matrix (*n*x*p*) consisting of *n* sample row vector *Xi* (s) we have:

Where *xij* (s) are partial-variables. Note that is called sample matrix.

Sample covariance is defined equivalently as below:

Where is *n*-dimension vector of ones.

Let where *In* is *n*x*n* identity matrix, we have another formula for computing sample covariance .

The expected value of is:

Due to:

And

We determine that sample covariance is unbiased estimate of population covariance *∑* because the expected value of equals *Σ*.

The sample correlation matrix is defined as below

Where is the sample partial-covariance between *kth* component and *lth* component of observation vectors and *sj* is the sample partial-standard-deviation of *jth* component of observation vectors. For convenience, let *x* and *y* represent the *kth* and *lth* components of observation vectors, the sample partial-correlation-coefficient is:

Where *xi* and *yi* denote *kth* and *lth* components of observation *Xi*, respectively.

Note that the sample partial-correlation-coefficient ranges in interval [–1*,* 1]. It is proved that *rkl* is unbiased estimate of theoretical correlation-coefficient *ρkl* = where *σkl, σk* and *σl* are theoretical covariance between *kth* and *lth* components, theoretical standard-deviation of *kth* component and theoretical standard-deviation of *lth* component, respectively. The definition of theoretical correlation-coefficient is described in previous section about multivariate distribution.

Suppose is the sample covariance matrix, the sample correlation-coefficient matrix is equivalently defined:

Where is diagonal matrix whose diagonal elements (*j, j*) are sample partial-variance of *jth* component.

In general, we often prefer unbiased estimates but biased estimates are used in some cases that it is impossible to draw unbiased estimates. If so, the best biased estimate is the one whose mean square error is smallest. The mean square error denoted *MSE*of estimate Θ is defined the expectation of square of deviation – *θ* as below:

Let and be biased estimates, the relative efficiency defined as below is used to choose which one is better.

If relative efficiency is less than 1, is more efficient than and otherwise. As aforementioned sample mean and sample covariance are unbiased estimates but there is a question that how to find out such estimates: unbiased estimates and efficient biased estimates. Methods to determine statistical estimates are described later.

### Parameter estimation methods

There are three methods to determine statistical estimates: moment method, maximum likelihood method and Bayesian method. Firstly, moment method is described because of its simplicity. The population or theoretical *kth* moment is defined as the expectation of *kth* power of random variable.

Where *Xk* is the *kth* power of *X*.

The *kth* sample moment is defined as below given a set of sample observations *X*1*, X*2*,…, Xm*.

For example, the first and second population moments are population mean *E*(*X*) *= μ* and population *E*(*X*2) = *E*(*XXT*) = *∑ + μμT*, respectively. The first and second sample moments are and . The basic idea of moment method is to equate population moment to sample moment. The estimates are solutions of such equations. For example, we set first and second population moments to be equal to first and second sample moments, respectively in order to find out the estimates of *∑* and *μ*.

is unbiased estimate of population mean *μ* and now we evaluate whether or not is unbiased estimate of population covariance *∑*.

Due to:

It implies that

So, is biased estimate of covariance *∑*. It means that moment method is not effective estimate method. Now we discuss a more popular method so-called maximum likelihood. The likelihood function is the joint probability which is the product of conditional probabilities of observations *Xi* (s) given the parameters. So likelihood function is considered as function of parameters. Let Θ = {*θ*1*, θ*2*,…, θk*} represent a set of parameters such population mean *μ*, population covariance *∑*, etc, the likelihood function *L*(Θ) is defined as below:

Let be sample estimate of Θ, we recognize that is extreme value at which *L*(Θ) gets maximal.

Because it is too difficult to work with the likelihood function in the form of product of condition probabilities, we should take the logarithm of *L*(Θ) so that the logarithm function converts the repeated multiplication to repeated addition. The natural logarithm of *L*(Θ) so-called log-likelihood is denoted *LogL*(Θ).

The essence of maximizing the likelihood function is to find the peak of the curve of *LogL*(Θ). This can be done by setting the first-order partial derivative of *LogL*(Θ) with respect to each parameter *θi* Θ to 0 and solving this equation to find out parameter *θi*. The reason is that the slope of the curve *LogL*(Θ) is equal to 0 at its peak. The number of equations corresponds with the number of parameters. For example, suppose observations *Xi* (s) conform normal distribution, we need to find out sample estimates of population mean *μ* and population covariance *∑*. The log-likelihood function is:

When is symmetric, we have:

It implies that the first-order partial derivative of *LogL*(Θ) with respect to *μ* is:

Now what we need to do is to first-order partial derivative of *LogL*(Θ) with respect to *∑*. We have two following partial derivatives with attention that *∑* is symmetric and invertible matrix.

It is easy to infer that the first-order partial derivative of *LogL*(Θ) with respect to *∑* is:

The estimates and are solutions of a set of equations which created by setting first-order partial derivatives of *LogL*(Θ) with respect to *μ* and *∑* to be zero.

Please pay attention that notation 0 denotes both zero vector and zero matrix.

We have:

Therefore the estimates of population mean and covariance are and , respectively. This result is similar to one from moment method when the estimate of covariance is biased estimate.

### Multivariate hypothesis testing

As aforementioned, hypothesis is defined as a statement about parameters of one or more populations. Hypothesis testing is the procedure that decides whether a hypothesis is accepted or rejected based on one or more sample statistic (s). Almost parametric hypothesis tests rely on two conditions:

* Sample observations conform normal distribution.
* Underlying distributions such as normal, chi-squared, student, *F*, etc are used to determine the critical region for tests.

The first condition is kept in multivariate tests but the second condition is the cause of testing complication because it is impossible to survey all of distributions such as normal, chi-squared, student, *F*, etc in multi-dimension data when sample points are vectors and some sample statistics are matrices. Statistical theory gives us an excellent method to eliminate multivariate distributions from testing process. In other words, likelihood function together with one-dimension distribution based on such likelihood function are used to calculate the critical region for tests instead of using underlying multivariate distributions. This method is called likelihood ratio test.

#### Likelihood ratio test

Suppose *p-*dimension data sample consists of *n* observation vectors *X*1*, X*2*,…, Xn*where *Xi* = {*xi*1, *xi*2,…, *xip*}. Note that *Xi* is identically distributed random vector and it can be called observation, data point, or sample point. In general case, suppose a parameter of the distribution of observations *Xi* (s) is *θ*, we test null hypothesis *H*0 that *θ* belongs to the null space which is the subspace of . Alternative hypothesis *H*1 is that *θ* belongs to the alternative space which is the subspace of Following is the general test case:

The simple test case is that the null hypothesis *H*0*: θ = θ*0 and *H*1: no constraint on *θ*; in other words, = {*θ*0} and = .

Note that if the distribution has many parameters, then *θ* belongs to set of parameters Θ = {*θ*1*, θ*2*,…, θk*}.

Let *L*0 and *L*1 be the likelihoods for null hypothesis *H*0 and alternative hypothesis *H*1, respectively.

Let and be the maxima of *L*0 and *L*1, respectively.

The likelihood ratio *R* is defined the ratio of null maximum to alternative maximum .

We take logarithm of *R* as below:

Note that and are maxima of *LogL*0 and *LogL*1 which, in turn, are logarithms of *L*0 and *L*1, respectively.

If null space is *r*-dimension space and alternative space is *q*-dimension space, it is proved that *–2Log*(*R*) is approximate to chi-square distribution *χ2* with *q – r* degrees of freedom.

It is easy to determine the rejection region of null hypothesis given significant level *α*; hence, *H*0 is rejected in flavor of *H*1 if

For example, given observations *Xi* (s) conform normal distribution *N*(*μ,∑*) where the theoretical covariance *∑* is known, the test case is:

We have = {*θ*0} and = and so the dimensions of and is 0 and *p*, respectively. The log-likelihoods *LogL*0 and *LogL*1 are:

When population covariance *∑* is known, the log-likelihood *LogL*0 is totally determined and the log-likelihood *LogL*1 gets maximal if *μ* =

.

Therefore, maxima and are totally determined.

It is easy to calculate the likelihood ratio and the test statistics –*2logR* in order to make decision whether or not reject the null hypothesis *H*0: *μ* = *μ*0, based on chi-square distribution with *p* degrees of freedom because of .

#### Test on sample correlation coefficient

Suppose *p-*dimension data sample consists of *n* observation vectors *X*1*, X*2*,…, Xn*where *Xi* = {*xi*1, *xi*2,…, *xip*}. For convenience, let *x* and *y* represent the *kth* and *lth* components of observation vectors, the sample partial-correlation-coefficient between two components *x* and *y* is:

Where *xi* and *yi* denote *kth* and *lth* components of observation *Xi*, respectively.

Let *ρxy* be the theoretical partial-correlation-coefficient. Note that *ρxy* ranges in interval [–1*,* 1]. If *ρxy* = 0, two components *x* and *y* are independent. When *rxy* is a statistic, we test on *ρxy* so as to determine whether or not two components *x* and *y* are independent. It means that we test on *H*0: *ρxy =* 0 in flavor of *H*1: no constraint on *ρxy*, based on statistic *rxy*. Although, it is very complicated to discover distribution of statistic *rxy*, there are some transformations on *rxy* can be applied. According to Fisher transformation, the statistic is approximate to normal distribution with mean and variance .

Where,

Given *H*0: *ρxy =* 0 and significant level *α*, *H*0 is rejected if the *z*-value, is greater than critical value defined by significant level *α*.

When sample size is small (*n* 25), Hotelling transformation is used:

Where *tanh* is hyperbolic tangent function.

Hotelling distribution *w\** is approximate to normal distribution with variance . Given *H*0: *ρxy =* 0 and significant level *α*, *H*0 is rejected if the *z*-value, is greater than critical value defined by significant level *α*.

Suppose *x* and *y* are normally distributed, *t*-distribution can be used to test on *H*0: *ρxy =* 0. The statistic conforms *t-*distribution with *n – 2* degrees of freedom.

We reject *H*0: *ρxy =* 0 if *|T| t*1– *α* /2*, n* – 2 with significant level *α*.

## Nonparametric hypothesis testing

This section is the brief survey of nonparametric hypothesis testing. It includes four main sub-sections about hypothesis testing, one additional section discussing goodness-of-fit and conclusion section.

* *Sign test* sub-section gives an overview of nonparametric testing, which begins with the test on sample median without assumption of normal distribution.
* *Signed-rank test* sub-section and *rank-sum test* sub-section concern improvements of sign test. The prominence of signed-rank test is to be able to test sample mean based on the assumption about symmetric distribution. Rank-sum test discards the task of assigning and counting plus signs and so it is the most effective method among ranking test methods.
* *Nonparametric ANOVA* sub-section discusses application of analysis of variance (ANOVA) in nonparametric model. ANOVA is useful to compare and evaluate various data samples at the same time.
* *Nonparametric goodness-fit-test* sub-section, an additional section, focuses on different hypothesis, which measure the distribution similarity between two samples. It determines whether two samples have the same distribution without concerning how the form of distribution is.

Note that in this report terms *sample* and *data sample* have the same meaning. A sample contains many data points. Each *data point* is also called an *observation*.

### Sign test

Nonparametric testing is used in case of without knowledge about sample distribution; concretely, there is no assumption of normality. The nonparametric testing begins with the test on sample *median*. If distribution is symmetric, median is identical to mean. Given the median is the data point at which the left side data and the right side data are of equal accumulate probability.

*P*(*D* <)= *P*(*D* >)= 0.5

If data is not large and there is no assumption about normality, the median is approximate to population mean. Given null hypothesis *H*0: = and alternative hypothesis *H*1: , the so-called sign test (Walpole, Myers, Myers, & Ye, 2012, pp. 656-660) is performed as below steps:

* Assigning plus signs to sample data points whose values are greater than and minus signs to ones whose values are less than . Note that values which equal are not considered. Plus signs and minus signs represent the right side and left side of , respectively.
* If the number of plus signs is nearly equal to the number of minus signs, then null hypothesis *H*0is true; otherwise *H*0 is false. In other words, that the proportion of plus signs is significantly different from 0.5 cause to rejecting *H*0 in flavor of *H*1.

The reason of *H*0 acceptance is that the probability that data points (or observations) fall in both left side and right side of are of equal value 0.5 and of course, it is asserted that is a real median. Note that terms *data point*, *sample point*, *sample value* and *observation* are identical.

In the case that alternative hypothesis *H*1: < , if the proportion of plus signs is less than 0.5 then rejecting *H*0 in flavor of *H*1. In the case that alternative hypothesis *H*1: > , if the proportion of plus signs is greater than 0.5 then rejecting *H*0 in flavor of *H*1. Now let *X* be the discrete random variable representing the number of plus signs and suppose that *X* conforms binomial distribution *B*(*X; n; p*) where *n* and *p* are the total number of sample data points and the probability that plus sign is assigned to a data point, respectively. Because the proportion of plus signs gets 0.5 when *H*0: is true, the parameter *p* is set to be 0.5. Given the distribution of plus signs is *B*(*X; n;* 0.5) and significant level *α* and let *x* be the instance of *X* where

There are three following tests (Walpole, Myers, Myers, & Ye, 2012, pp. 657-660):

* *H*0: = and *H*1: : In case of *x < n*/2, if 2*P*(*X x*) *< α*  then rejecting *H*0. In case of *x > n*/2, if 2*P*(*X x*) *< α* then rejecting *H0*. This test belongs to two-sided test family.
* *H*0: = and *H*1: < : if *P*(*X x*) *< α* then rejecting *H*0. This test belongs to one-sided test family.
* *H*0: = and *H*1: > : if *P*(*X x*) *< α*  then rejecting *H*0. This test belongs to one-sided test family.

Note that *P*(…) is accumulated probability of binomial distribution *B*(*X; n;* 0.5), for example,

In case that *n* is large enough, for instance *n* > 30, *B*(*X; n;* 0.5) is approximate to standard normal distribution *N*(*Z;* 0*;* 1) where

Let *z* be the instance of *Z*, there are three following tests:

* *H*0: and *H*1: : if *|z| > zα*/2 then rejecting *H*0 where *zα*/2 is 100*α*/2 percentage point of standard normal distribution.
* *H*0: and *H*1: : if *z <* –*zα*/2 then rejecting *H*0.
* *H*0: = and *H*1: : if *z > zα*/2 then rejecting *H*0.

In case of pair-test *H*0: – = *d*0 which we need to know how much median shifts from other one , sign test is applied in similar way with a little bit of change. If *d*0 *=* 0, *H*0 indicates whether equals . We compute all deviations between two samples *X* and *Y* where is sample median of *X* and is sample median of *Y*. Let *di = xi – yi* be the deviation between *x* *Y* and *y* *Y*. Plus signs (minus signs) are assigned to *di* (s) which are greater (less) than *d*0. Now signed test is applied into such plus signs and minus signs by discussed method.

### Signed-rank test

Sign test focuses on whether or not the observations are different from null hypothesis but not considers the magnitude of such difference. Wilcoxon signed-rank test (Walpole, Myers, Myers, & Ye, 2012, pp. 660-663) based on assumption of symmetric and continuous distribution considers both difference and how much difference is. The median is identical to the mean *μ* according to symmetric assumption. It includes four following steps (Walpole, Myers, Myers, & Ye, 2012, pp. 660-663):

1. Calculating all deviations between data points and *μ*0, we have *D =* {*d*1*, d*2*,…, dn*} where *di = xi – μ*0 and *di* 0.Note that data point *xi* is instance of random variable *X*.
2. Assigning a rank *ri* to each deviation *di* without regard to sign, for instance, rank value 1 and rank value *n* to be assigned to smallest and largest **absolute** deviation (without sign), respectively. If two or more absolute deviations have the same value, these deviations are assigned by average rank. For example, if 3*rd*, 4*th* and 5*th* deviations get the same value, they receive the same rank (3*+*4*+*5) / 3 = 4. We have a set of ranks *R =* {*r*1*, r*2*,…, rn*} where *ri* is the rank of *di*.
3. Let *w+* and *w*– be the sum of ranks whose corresponding deviations are positive and negative, respectively. We have *w+ =*  and *w*– *=*  and *w = min*(*w+, w*–). Note that *w* is the minimum value between *w+* and *w*–.
4. In flavor of *H*1: *μ*< *μ*0, *H*0 is rejected if *w+* is sufficiently small. In flavor of *H*1: *μ* > *μ*0, *H*0 is rejected if *w*– is sufficiently small. In case of two-sided test *H*1: *μ μ*0, *H*0 is rejected if *w* is sufficiently small. The concept “sufficiently small” is defined via thresholds or pre-computed critical values, see [Walpole, Myers, Myers, Ye 2012, pp. 759] for critical values. The value *w+, w–* or *w* is sufficiently small if it is smaller than a certain critical value with respect to significant level *α*.

In case of pair test *H*0: *μ*1– *μ*2 *= d*0, the deviation *di* in step 1 is calculated based *d*0 and two samples *X* and *Y*, so *di = xi – yi* *– d*0where *x* *Y* and *y* *Y*. Note that *μ*1 and *μ*2 are taken from *X* and *Y*, respectively. Steps 2*,* 3*,* and4 are performed in similar way.

Let *W+* be random variables of *w+*. If *n 15* then *W+* approaches normal distribution with mean

and variance

We can normalize *W+* so as to define critical region via which is 100(1 – *α*) percentage point *zα* of normal standard distribution.

### Rank-sum test

Rank-sum test (Walpole, Myers, Myers, & Ye, 2012, pp. 665-667) is a variant of signed-rank test. Suppose there are two samples *X* = {*x*1*, x*2*,…,* } and *Y =* {*y*1*, y*2*,…,* } and the null hypothesis is specified as *H*0: *μ*1= *μ*2 where *μ*1 and *μ*2are taken from *X* and *Y*, respectively. We assign ranks to such *n*1 *+ n*2 data points according to their values, for instance, rank value 1 and rank value *n*1 *+ n*2 to be assigned to smallest and largest sample value. If two or more data points have the same value, these points are assigned by average rank. For example, if 3*rd*, 4*th* and 5*th* data points get the same value, they receive the same rank (3*+*4*+*5) / 3 = 4*.* Let *R =* {*r*1*, r*2*,…,* } be the set of these ranks. Let *w*1 and *w*2 be the sum of ranks corresponding to *n*1 data points in *X* and *n*2 data points in *Y*, respectively.

and

Where *ri* is a rank of a data point in the set *X Y* and *ri =* .

We have . There are three following tests (Walpole, Myers, Myers, & Ye, 2012, pp. 665-667):

* Rejecting *H*0 in flavor of alternative *H*1: *μ*1< *μ*2 if *w*1 is sufficiently small.
* Rejecting *H*0 in flavor of alternative *H*1: *μ*1> *μ*2 if *w*2 is sufficiently small.
* In case of two-sided test with *H*1: *μ*1 *μ*2 if the minimum of *w*1 and *w*2 is sufficiently small then rejecting *H*1.

Rank-sum test has two advantages in comparison of signed-rank test:

* There is no need to calculate deviations among samples and to count the number of plus signs and minus signs.
* Samples can has different number of data points, for instance, |*X*| = *n1* *n2* = |*Y*|.

Setting *u*1 *= w*1 – and *u*2 *= w*1 – and suppose that *u*1 and *u*2 are instances of random variables *U*1 and *U*2, respectively. If both *n*1and *n*2are greater than8, variable *U*1 (or *U*2) is approximate to normal distribution with mean

and variance

We can normalize *U*1 (*U*2) so as to define critical region via which is 100(1 – *α*) percentage point *zα* of normal standard distribution

### Nonparametric ANOVA

In many applications, we process various samples (*X*, *Y*, *Z*, etc.) where each sample is a set of observations (data points) which relate to a concrete method, a way or an approach that creates or produces these observations. Such concrete method is called **treatment**. In other words, we consider a matrix of observations and each row represents a mono-sample attached to a treatment, for instance, *X* or *Y* or *Z*, etc. For convenience, matrix of observations is call multi-sample or sample, in short. Treatments are grouped into categories which are called **factors**. If sample has only one factor, it is single-factor sample; otherwise, it is called several-factor sample. Following table is an example of single-factor sample.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Treatment 1 | *y11* | *y12* | *y13* | *=*( *y11 + y12 + y13*) / 3 |
| Treatment 2 | *y21* | *y22* | *y23* | *=*( *y21 + y22 + y23*) / 3 |
| Treatment 3 | *y31* | *y32* | *y33* | *=*( *y31 + y32 + y33*) / 3 |
|  |  |  |  | *=*( *+ +* ) / 3 |

Let *Yij* be the random variable representing *jth* data point of *ith* treatment.

Where *μ* so-call overall mean is the mean over whole sample, called treatment effect denotes the parameter of *ith* treatment and denotes the random error. There is the assumption that random error is independently distributed and conforms normal distribution; moreover, it has mean 0 and variance *σ*2. Let *μi = μ + τi* be the treatment mean of *ith*treatment. The objective of analysis of variance (ANOVA) (Montgomery & Runger, 2003, pp. 468-490) is to analyze statistics about treatment mean, treatment effect, random error so as to take out conclusions about such statistics. Basically, ANOVA focuses on characteristics relating to deviation, variability, sum of squares, mean square, etc. An typical of ANOVA is to test whether *k* treatment means *μ1, μ2,…, μk* are equal; it means that we test the following hypotheses:

*H*0*: μ*1 *= μ*2 *= …= μk*

*H*1: *μ*1 *μ*2 *… μk*

Due to *μi = μ + τi* , this test is re-written:

*H*0*: τ*1 *= τ*2 *= …= τk* = 0

*H*1: *τi* 0 for at least one treatment

If *H*0 is true, treatments have no effect on whole sample. Let *yij* be the instance of random variable *Yij*. Let *yi, , y* and be the sum of observations of treatment *i*, the average of observations of treatment *i*, the sum of whole observations and the average of whole observations.

, , ,

Where *k* is the number of treatments, *ni* is the number of observations under treatment *i* and *N = n*1 *+ n*2 *+… + nk* is the total number of observations. Let *SST*, *SSTreatment* and *SSE* (Montgomery & Runger, 2003, pp. 474-475) be the total sum of squares, treatment sum of squares and error sum of squares. Pleasepay attention to *SST*, *SSTreatment* and *SSE* because they are main research objects in ANOVA. We have (Montgomery & Runger, 2003, p. 475):

Following is the sum of squares identity (Montgomery & Runger, 2003, p. 475):

*SST = SSTreatment + SSE*

Treatment sum of squares *SSTreatment* is very important because it reflects treatment effects *τi* (s) and treatment means *μi* (s). The expected values of treatment sum of squares and error sum of squares are computed as below (Montgomery & Runger, 2003, p. 474):

*SST* and *SSTreatment* and *SSE* have *N –* 1 and *k –* 1 degrees of freedom, respective because there are *N* observations over whole sample and *k* treatments. So *SSE* has *N – k =* (*N –* 1) *–* (*k –* 1) due to *SSE = SST – SSTreatment*. Based on degrees of freedom, treatment mean square *MSTreatment* and error mean square *MSE* is determined as below (Montgomery & Runger, 2003, pp. 474-475):

If null hypothesis *H*0*: τ*1 *= τ*2 *= …= τk* = 0 is true, *MSTreatment* is an unbiased estimate of variance σ2 due to:

Moreover *MSE* is always an unbiased estimate of variance *σ*2 due to:

So *MSTreatment* and *MSE* conform chi-square distribution and the ratio of *MSTreatment* to *MSE* conforms *F-*distribution with *k –* 1 and *N – k* degrees of freedom. We have (Montgomery & Runger, 2003, p. 475):

Hypothesis *H*0*: τ*1 *= τ*2 *= …= τk* = 0 is rejected if the ratio *F*0 > *fα, k–*1*, N–k* where *fα, k–*1*, N–k* is the 100(1– *α*) percentage point of *F-*distribution with *k –* 1 and *N – k* degrees of freedom.

We have already discussed about parametric ANOVA with normality assumption, now nonparametric ANOVA is the next topic. Nonparametric ANOVA has no normality assumption of random error but the independence of random error is required. The Kruskal-Wallis [Montgomery, Runger 2003] [Walpole, Myers, Myers, Ye 2012] test is a popular nonparametric test. Suppose treatment *i* has *ni* observations and there are *k* treatment, let *N = n*1 *+ n*2 *+…+ nk* be the total of observations. Kruskal-Wallis test assigns ranks to such *N* observations according to their values, for instance, rank value 1 and rank value *N* to be assigned to smallest and largest sample value. If two or more observations have the same value, these observations are assigned by average rank. For example, if 3*rd*, 4*th* and 5*th* observations get the same value, they receive the same rank (3*+*4*+*5) / 3 = 4*.* Let *Rij* be the rank of observation *Yij*. If null hypothesis *H*0: is true, which means that all treatments have the same mean, then ranks spread over all treatments equally. In other words, the expected value of *Rij* (s) is nearly equal to the mid-point of *N* ranks, so we have (Montgomery & Runger, 2003, p. 589):

*E*(*Rij*) *=* (*N +* 1)/2

Let = be average rank of treatment *i*, the expected value of is determined as below (Montgomery & Runger, 2003, p. 589):

If the null hypothesis *H*0: is true, the average rank does not shift from its expected value (*N* + 1) / 2 much. The difference between and its expected value (*N* + 1) / 2 is determined by following statistic (Montgomery & Runger, 2003, p. 589):

This formula is transformed into more practical format as below (Montgomery & Runger, 2003, p. 590):

Where is the sum of ranks under treatment *i*. It is proved that statistic *K* approaches chi-square distribution with *k –* 1 degrees of freedom where *k* is the number of treatments. Null hypothesis *H*0: is rejected in flavor of alternative hypothesis *H*1: if *K >* .

### Nonparametric goodness-fit-test

Goodness-fit-test is the test that determines whether a sample conforms specified distribution or whether two samples have the same distribution. Although Kolmogorov–Smirnov goodness-fit-test being a kind of nonparametric testing does not consider the sample distribution, it is based on the definition of Kolmogorov distribution. Kolmogorov distribution is continuous distribution whose accumulative distribution function is defined as below (Wikipedia, 2016):

The critical value *Kα* at significant level *α* is 100(1 *– α*) percentage point satisfying equation:

Nonparametric Kolmogorov–Smirnov test is to determine whether two samples have the same distribution regardless of the underlying distribution. Given *X =* {*x*1*, x*2*,…, xn*} and *Y =* {*y*1*, y*2*,…, yn*} are two testing samples, the null hypothesis *H*0 is that *X* and *Y* have the same distribution. Let *FX* and *FY* be the empirical distribution functions of *X* and *Y*, respectively. Note that empirical distribution function is accumulative function which increases gradually according to the order of values.

Let *D* be the maximum absolute deviation between *FX* and *FY* over whole samples *X* and *Y*.

It is easy to recognize that the process to find out *D* is iterative process browsing all pairs of observation (*xi, yi*) *X Y*. It is proved that *D* conforms *K* distribution. Therefore, the null hypothesis *H*0 is rejected at significant level *α* if *D > Kα*.

### Conclusion

Now we had a general and detailed point of view about nonparametric testing. We can draw two main comments from research over this domain:

* Firstly, nonparametric model is less efficient than parametric model because it lacks valuable information under sample when it has no knowledge about the distribution. All properties of distribution such as mean, variance, standard deviation, median, mode, skewness, kurtosis, etc are essential information of which nonparametric model does not take advantages. However, nonparametric testing is very useful and appropriate to cases that knowledge of distribution cannot be extracted or sample does not conform normal distribution. In case that underlying distribution is ignored and nonparametric testing is the best choice. Therefore, we conclude that the most important thing is to choose appropriate model (parametric or nonparametric) which is adaptive to testing situation and testing requirement.
* Secondly, nonparametric model is often based on ranking. Ranking process aims to transform origin sample into simpler sample so-called ranking sample. Ranking sample is the set of ranks; thus, each rank is assigned to respective observation from origin sample. Because nonparametric model does not know valuable information of origin sample such as mean, variance, standard deviation; it will exploit ranking sample to discover such valuable information. Therefore, nonparametric testing, in turn, applies parametric methods into the ranking sample. Concretely, nonparametric testing assumes that statistic (s) on ranking sample conform some pre-defined distributions. For example, sign test assumes that the number of plus signs in ranking data conforms binominal distribution, signed-rank test and sum-rank test apply Wilcoxon distribution into ranking data and nonparametric goodness-fit-test is based on Kolmogorov distribution. We conclude that parametric testing and nonparametric testing have a strongly mutual relationship and so, we should take advantages of both of them.

# Multivariate parametric hypothesis testing

# Hypothesis testing within incomplete multivariate data

Multivariate hypothesis testing becomes more and more necessary when data is in the process of changing from scalar and univariate format to multivariate format, especially financial and biological data is often constituted of *n-*dimension vectors. As aforementioned, likelihood ratio test is the best method that applies the test on mean of multivariate sample with known or unknown covariance matrix but it is impossible to use likelihood ratio test in case of incomplete data when the data incompletion gets popular because of many reasons in reality. Therefore, this research proposes a new approach that gives an ability to apply likelihood ratio test into incomplete data. Instead of replacing missing values in incomplete sample by estimated values, this approach classifies incomplete sample into groups and each group is represented by a potential or partial distribution. All partial distributions are unified into a mixture model which is optimized via expectation maximization (EM) algorithm. Finally, likelihood ratio test is performed on mixture model instead of incomplete sample. This research provides a thorough description of proposed approach and mathematical proof that is necessary to such approach. The comparison of mixture model approach and filling missing values approach is also discussed in this research.

## Mixture model test on incomplete data

Suppose *p-*dimension data sample consists of *n* observation vectors *X*1*, X*2*,…, Xn*where *Xi* = {*xi*1, *xi*2,…, *xip*}. Thus,  *=* { *X*1*, X*2*,…, Xn*} compose a matrix whose each row is observation *Xi*. In case of incomplete data, is sparse matrix and observation *Xi* is not always complete vector when it can lack some components *xij* (s). Likelihood ratio test cannot apply into incomplete sample as because it is impossible to calculate statistics such as mean and variance in terms of incomplete sample. This research tries to overcome this drawback by discovering potential probability distributions under observations *Xi* (s) regardless of data incompletion. Suppose we test on mean of normal distribution when variance known or unknown, so *Xi* (s) conforms *N*(*μ, ∑*) and the null hypothesis *H*0*: μ = μ*0 is in flavor of alternative hypothesis *H*1: no constraint on *μ*. Based on such potential distributions, the maximum likelihood of these hypotheses is determined.

Firstly, we estimate the number of potential probability distributions under observations *Xi* (s). Observations are classified into *k* classes following two conditions below.

1. Class *c*1 represents observations whose components *xij* (s) are completed. Note that class *c*1 may not exist.
2. Classes *c*2*, c*3*,…, ck* represent observations lacking the same components *xij* (s). For example, we have four observations *X*1 = {*x*11*=*1*, x*12*=*2*, x*13(*empty*)*, x*14(*empty*)}, *X*2 *=* {*x*21*=*2*, x*22*=*3*, x*23(*empty*)*, x*24(*empty*)}*, X*3 *=* {*x*31(*empty*)*, x*32(*empty*)*, x*33*=*1*, x*34 *=*2}and *X*4 *=* {*x*41(*empty*)*, x*42(*empty*)*, x*43*=*2*, x*44*=*3}. Thus, *X*1 and *X*2 belong to the same class (class 1); *X*3 and *X*4 belong to the same class (class 2) according to this condition. Note that empty value is considered as missing value. Two or many classes can overlap together. For example, if *X*5 = {*x*11*=*3*, x*12*=*(*empty*)*, x*13=(*empty*)*, x*14(*empty*)} then *X*5 belongs to different class – class 3, thus, class 1 and class 3 are overlapped.

The number of potential probability distributions is initialized to be *k*. Note that *k* should be much smaller than the size *n* of sample. Note that potential probability distribution is also called partial probability distribution. Let *pi* be the potential probability distributions corresponding to class *i*. Suppose *pi* conforms normal distribution *N*(*μi*, *∑i*) with mean *μi* and variance *∑i*. The probabilistic mixture model (Bilmes, 1998, p. 3) is defined as below.

Where,

Where *pj* is the potential distribution and *xi* is instance of variable *Xi*. Let Θ = {*α*1, *α*2,*…*, *αk*, *θ*1, *θ*2,*…*, *θk*} and *θj* = {*μj, ∑j*} be probabilistic parameters of sample and potential distribution *pj*, respectively. Formula (III.1) indicates that the distribution of data point *xi* is constituted of potential (or partial) distributions *pj* (s). Each potential distribution *pj* is weighted by the weight *αj* such that . The weight *αj* is the probability of *pj* if we consider *pj* as random variable. The weights *αj* (s) are learned and updated from sample, which will be discussed later. The likelihood function of is:

We take the logarithm of likelihood function so as to convert the repeated multiplication into repeated addition, so the log-likelihood function of is (Bilmes, 1998, p. 3):

Let *Y*1*, Y*2*,…, Yn* be variables indicating that data point *Xi* comes from which potential distribution. The value of each *Yi* ranges in {1*,* 2*,* 3*,…, n*}. Concretely, if *Yi = yi* then data point *Xi* conforms distribution . When  *=* {*X1, X2,…, Xn*} is incomplete data,  *=* {*Y1, Y2,…, Yn*} is observational data that let us exploit . The probability of *Yi = yi* is the prior probability of which indicates that how weighted data point *xi* comes from distribution , so we have *P*(*yi*) = . Similarly, the conditional probability of data point *xi* given *yi* is the partial probability *pj* given *xi*, so we have *P*(*xi | yi*) = . Let be the complete data, the log-likelihood function is re-written as below (Bilmes, 1998, p. 3).

Where,

Our goal is to find out the optimal parameter Θ*\** = {*α*1*\*, α*2*\*,…, αk\*, θ*1*\*, θ*2*\*,…, θk\**} that maximizes the log-likelihood function in (III.2). Expectation maximization (EM) algorithm being the iterative process is applied in order to find out Θ*\**. Let be the optimal parameter at *tth* iteration. Note that *θjt =* {*μjt, Σjt*} is the optimal parameter of partial probability *pj* at *tth* iteration. In general, starting with initial estimate Θ0, each iteration in EM algorithm has two steps (Sean, 2009, p. 8):

* E-step: computing the conditional expectation based on the previous estimate Θ*t*.
* M-step: finding out current estimate Θ*t+*1 *=* Θ*\** that maximizes such conditional expectation. Note that Θ*t+*1 is reserved for next iteration.

EM algorithm stops when it meets the terminating condition, for example, the difference of previous estimate Θ*t* and current estimate Θ*t+*1 is smaller than some pre-defined threshold *ε*, namely, |Θ*t+*1 *–* Θ*t*| < *ε*. In E-step, the conditional expectation *Q*(Θ*,* Θ*t*) is determined as below (Bilmes, 1998, p. 4).

Note that *Q*(Θ*,* Θ*t*) is the function of variable Θ and the value Θ*t* is known estimate in previous iteration. The conditional expectation *Q*(Θ*,* Θ*t*) sums over all possible instances of *=* {*Y1, Y2,…, Yn*} and so *Ψ* denotes all combinational values over . Now we need to specify the probability of *y* given and Θ*t*, *P*(*y | ,* Θ*t*). Note that *y =* {*y1, y2,…, yn*} is instance of observation variable data  *=* {*Y1, Y2,…, Yn*}. Applying Bayesian rule, we have:

Suppose *y1, y2,…* and *yn* are mutually independent, the probability of *y* given is following product (Bilmes, 1998, p. 4).

Consequently, the conditional expectation *Q*(Θ*,* Θ*t*) is expended as below.

Note that *P*(*yi = c | xi*) is computed following formula (III.4). Because formula (III.5) has two parameters *αc* and *θc =* {*μc, ∑c*}, the conditional expectation *Q*(Θ*,* Θ*t*) is split into two parts as below.

Let

It is easy to recognize that both *G*(*αc*) and *H*(*θc*) are less than or equal to 0 because of . Hence, the expectation *Q*(Θ*,* Θ*t*) is maximal if and only if both *G*(*αc*) and *H*(*θc*) are minimal. In other words, we find out extreme points *αc\** and *θc\** = {*μc*\**, ∑c\**} that minimize both *G*(*αc*) and *H*(*μc*,*∑c*) when *Q*(Θ*,* Θ*t*) is considered as function of variables *αc* and *θc* = { *μc*,*∑c* }.

Applying Lagrange function into *G*(*αc*) with constraint , extreme point *αc\** is the solution of the equation formed by setting first-order derivative of sum of *G*(*αc*) and Lagrange constraint to be zero.

Summing equation (III.7) over *k* classes {1*,* 2*,…, k*}, we have (Bilmes, 1998, p. 5):

Substituting *λ = –n* into equation (III.7), extreme point *αc\** is totally determined.

Where *P*(*yi = c | xi*) is calculated following equation (III.4).

Now we find out another extreme point *θc\** = {*μc*\**, ∑c\**}. Suppose potential distribution *pc*(*xi |θc*) is normal, it is expended as below.

It implies that

Where *p* is the dimension of sample space .

The first-order partial derivative of *H*(*μc*, *∑c*) with respect to *μc* is:

The optimal *μc*\* maximizing *H*(*μc*, *∑c*) is the solution of equation created by setting partial derivatives of *H*(*μc*, *∑c*) with regard to mean *μc* to be 0. Note that 0 denotes zero vector.

The first-order partial derivative of *H*(*μc*, *∑c*) with respect to *∑c* is:

Due to:

And

Because author (Bilmes, 1998, p. 5) mentioned:

Where tr(*A*) is trace operator which takes sum of diagonal elements of matrix .

It implies (Nguyen, 2015, p. 45)

Where *∑c* is symmetric and invertible matrix.

The optimal *∑c*\* maximizing *H*(*μc*, *∑c*) is the solution of equation created by setting partial derivatives of *H*(*μc*, *∑c*) with regard to *∑c* to zero. Note that 0 denotes zero matrix, we have:

In general, the optimal parameters *αc\** and *θc\** = {*μc*\**, Σc\**} for each potential (partial) probability *pc* form a following triple. Note that *P*(*yi = c | xi*) is computed following formula (III.4).

Of course, we have *k* groups (*α*1*\** , *θ*1*\**), (*α*2*\** , *θ*2*\**),…, and (*αk\** , *θk\**) like (III.8) and so the global optimal parameter Θ*\** = {*α*1*\*, α*2*\*,…, αk\*, θ*1*\*, θ*2*\*,…, θk\**} in M-step of EM algorithm is totally specified. The basic idea is to divide the global optimal parameter Θ*\** into *k* groups of partial parameters (*αj\** , *θj\**) and find out such partial parameters in particular.

Back the test on mean of normal distribution when the null hypothesis *H*0*: μ = μ*0 is in flavor of alternative hypothesis *H*1: no constraint on *μ*. The likelihood ratio *R*() is defined as the ratio of the maximum likelihood of null hypothesis to the maximum likelihood of alternative hypothesis.

Where and are maximum likelihoods of null hypothesis and alternative hypothesis, respectively.

This research proposes a 4-step testing process based on mixture model in order to test on mean of normal distribution in case of incomplete data.

1. Specifying *k* classes and *k* respective partial probabilities *pc*(*xc | θc*) with and *θc* = {*μc, ∑c*}. The weights (s) of partial probabilities *pc* (s) are initialized by the ratio of the number *Xi*(s) belonging to class *c* to *n*, the number of total *Xi* (s).
   1. Class *c*1 represents observations whose components *xij* (s) are completed. Note that class *c*1 may not exist.
   2. Classes *c*2*, c*3*,…, ck* represent observations lacking the same components *xij* (s).
2. Specifying the likelihood functions of null hypothesis and alternative hypothesis such as *L0*() and *L1*().
   1. where parameter *θc* = {*μc*, *∑c*} is constant and is assigned by sample mean *μ0* and known covariance matrix *∑* (or sample covariance matrix ).
   2. .
3. Applying EM algorithm into finding out optimal parameters *αc\** and *θc\** = {*μc*\**, ∑c\**} for each potential (partial) probability *pc*.
   1. Optimal parameters *αc\** with respect to null hypothesis: . Note that *θc* = {*μc, ∑c*} is constant in .
   2. Optimal parameters *αc\** and *θc\** = {*μc*\**, ∑c\**} with respect to alternative hypothesis:
4. Substituting *αc\** and *θc\** = {*μc*\**, ∑c\**} into likelihood functions *L*0() and *L*1() so that the likelihood ratio is totally determined. It is proved that *–2log*(*R*) is approximate to chi-square distribution *χ2* with *p* degrees of freedom when population covariance *∑* is known; so *H*0 is rejected in flavor of *H*1 if *–2log*(*R*) > *χ2*1 *– α, p* with significant level *α*.

Please pay attention to step 1 and step 3 because it is slightly complicated. Suppose data sample is *p-*dimension vector space , which contains *n* observation vectors *X*1*, X*2*,…, Xn*where *Xi* = {*xi*1, *xi*2,…, *xip*}. Thus,  *=* {*X*1*, X*2*,…, Xn*} compose a matrix whose each row is observation *Xi*. Suppose *Xi* (s) conforms normal distribution *N*(*μ, ∑*). If class *c* is composed of *u* non-empty components where *cu* = , then the mean *μc* contains only *c*1*th, c*2*th,…, cuth* components which correspond with *c*1*th, c*2*th,…, cuth* columns in matrix . Similarly, the covariance matrix *∑c* contains only variances among *c*1*th, c*2*th,…, cuth* components which correspond with *c*1*th, c*2*th,…, cuth* columns in matrix . All arithmetical operators in step 3 are performed based on such *u* non-empty components . This important thing is the tip of this research. The mean *μc* in the likelihood function of null hypothesis is constant and is assigned by *c*1*th, c*2*th,…, cuth* components of sample mean *μ*0*.* The covariance matrix *∑c* in the likelihood function of null hypothesis is constant and is assigned by *c*1*th, c*2*th,…, cuth* components of population covariance *∑* or sample covariance in case of unknown *∑*. It means that it is not necessary to computed the optimal parameters *μc\** and *∑c\** with respect to likelihood function of null hypothesis *L*0().

For example, sample has four observations *X*1 = {*x*11*=*1*, x*12*=*2*, x*13(*empty*)*, x*14(*empty*)}, *X*2 *=* {*x*21*=*2*, x*22*=*3*, x*23(*empty*)*, x*24(*empty*)}*, X*3 *=* {*x*31(*empty*)*, x*32(*empty*)*, x*33*=*1*, x*34 *=*2}and *X*4 *=* {*x*41(*empty*)*, x*42(*empty*)*, x*43*=*2*, x*44*=*3}. These observations formed a 4x4 matrix shown in following table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *X*1 | *x*11 = 1 | *x*12 = 2 | *x*13 = ? | *x*14 = ? |
| *X*2 | *x*21 = 2 | *x*22 = 3 | *x*23 = ? | *x*24 = ? |
| *X*3 | *x*31 = ? | *x*32 = ? | *x*33 = 1 | *x*34 = 2 |
| *X*4 | *x*41 = ? | *x*42 = ? | *x*43 = 2 | *x*44 = 3 |

**Table III.1.** An example of incomplete sample where question mark (?) denotes missing value (or empty value).

Suppose we test on mean of normal distribution and so *Xi* (s) conforms normal distribution *N*(*μ, ∑*) and the null hypothesis *H*0*: μ = μ0* = (0*,* 0*,* 0*,* 0)*T* and *H*1: no constraint on *μ*. Suppose variance is known, *∑* = . Sample is classified into two classes such as class 1 containing *X*1, *X*2 and class 2 containing *X*3, *X*4. Let *p*1 and *p*2 are partial probabilities attaching to class 1 and class 2, respectively. Both *p*1 and *p*2 conform normal distribution, *p*1 *N*(*μ*1*, ∑*1) and *p*2 *N(μ*2*, ∑*2). Let  *=* {*Y*1, *Y*2, *Y*3, *Y*4} be the set of observations where *Yi* indicates *Xi* belongs to which class (1 or 2). Because *X*1, *X*2 belong to class 1 and *X*3, *X*4 belong to class 2, we have  *=* {*Y*1 *=* 1, *Y*2 *=* 1, *Y*3 *=* 2, *Y*4 *=* 2}*, p*1(*x*1) *=* 0.5, *p*1(*x*2) *=* 0.5, *p*1(*x*3) *=* 0, *p*1(*x*4) *=* 0, *p*2(*x*1) *=* 0, *p*2(*x*2) *=* 0*, p*2(*x*3) *=* 0.5and *p*2(*x*4) *=* 0.5.

When class 1 lacks columns (components) 3 and 4 and relates only *X*1 and *X*2, the mean *μ*1 receives two first components of *μ* and the covariance *∑*1 containing variances between *X*1 and *X*2 is the sub-matrix of *∑*. Similarly, the mean *μ*2 receives two last components of *μ* and the covariance *∑*2 is the sub-matrix of *∑* which contains variances between *X*3 and *X*4. We have:

Let *α*1 and *α*1 be the weights of partial probabilities *p*1 and *p*2, respectively. Hence, *αi* is initialized as the number of *Xi* (s) belonging to class *i*. We have:

At the first iteration of EM algorithm, the probability of *y* given *X* denoted *P*( *yi | xi*) is calculated according to formula (4), for example:

In the similar way, we have:

* *P*(*y*1=1| *x*1) = 1, *P*(*y*1=1| *x*2)=1, *P*(*y*1=1| *x*3)=0, *P*(*y*1=1| *x*4)=0, *P*(*y*1=2| *x*1) = 0, *P*(*y*1=2| *x*2)=0, *P*(*y*1=2| *x*3)=1, *P*(*y*1=2| *x*4)=1.
* *P*(*y*2=1| *x*1) = 1, *P*(*y*2=1| *x*2)=1, *P*(*y*2=1| *x*3)=0, *P*(*y*2=1| *x*4)=0, *P*(*y*2=2| *x*1) = 0, *P*(*y*2=2| *x*2)=0, *P*(*y*2=2| *x*3)=1, *P*(*y*2=2| *x*4)=1.
* *P*(*y*3=1| *x*1) = 1, *P*(*y*3=1| *x*2)=1, *P*(*y*3=1| *x*3)=0, *P*(*y*3=1| *x*4)=0, *P*(*y*3=2| *x*1) = 0, *P*(*y*3=2| *x*2)= *0*, *P*(*y*3=2| *x*3) *=* 1, *P*(*y*3 = 2| *x*4) *=* 1.
* *P*(*y*4 = 1| *x*1) *= 1*, *P*(*y*4 = 1 | *x*2)= 1, *P*(*y*4 = 1 | *x*3) = 0, *P*(*y*4 = 1 | *x*4) = 0, *P*(*y*4 = 2 | *x*1) = 0, *P*(*y*4 = 2 | *x*2) = 0, *P*(*y*4 = 2 | *x*3) = 1, *P*(*y*4 = 2 | *x*4) = 1.

Parameters *μ*1, *∑*1, *μ*2, *∑*2*, α*1 and *α*2 are re-calculated according to formula (III.8).

Where *μ*1*\**, *∑*1*\**, *μ*2*\**, *∑*2*\*, α*1*\** and *α*2*\** are optimal parameters that maximizing log-likelihood function. Because the deviation between *α*1*\** and *α*10 is zero and the deviation between *α*2*\** and *α*20is zero, algorithm is stopped at the first iteration. Let *R*() be the likelihood ratio.

Where *L0\**() and *L1\**() are maximum likelihoods of null hypothesis and alternative hypothesis, respective. We have:

Where *LogL0\**(*X*) and *LogL1\**(*X*) are maximum likelihoods of null hypothesis and alternative hypothesis, respective. We have:

Thus, hypothesis *H*0: *μ*0 = (0*,* 0*,* 0*,* 0)*T* cannot be rejected in flavor of *H*1: no constraint at significant level *0.05*  because *–2logR = 0 < 9.49 = χ20.05,4*.

## Mixture model vs. filling missing values

Filling missing values method is to try to estimate missing values (or empty values) so as to transform incomplete data into complete data and after that, hypothesis testing is done on complete data as usual. We discuss some methods of filling missing values and compare them with mixture model. Suppose data sample is *p-*dimension vector space, which contains *n* observation vectors *X*1*, X*2*,…, Xn* where *Xi* = {*xi1*, *xi2*,…, *xip*}. Thus,  *=* { *X1, X2,…, Xn*} composes a matrix whose each row is observation *Xi*. In case of incomplete data, *X* is sparse matrix and observation *Xi* is not always complete vector when it can lack some components *xij* (s). Following table is an example of sample .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *v* = 1 | *v* = 2 | *v* = 3 | *v* = 4 |
| *u* = 1 | *x*11 = 1 | *x*12 = 2 | *x*13 = ? | *x*14 = ? |
| *u* = 2 | *x*21 = 2 | *x*22 = 3 | *x*23 = ? | *x*24 = ? |
| *u* = 3 | *x*31 = ? | *x*32 = ? | *x*33 = 1 | *x*34 = 2 |
| *u* = 4 | *x*41 = ? | *x*42 = ? | *x*43 = 2 | *x*44 = 3 |

**Table III.2.** Incomplete sample where question mark (?) denotes missing value (or empty value), *u* denotes row index and *v* denotes column index.

In the easiest way, if each column is considered as random variable, the missing value will be estimated as the mean of column vector. This way is called average method.

Where *K* is the number of non-empty values.

According to average method, we have *x*31 = *x*41 = (1+2) / 2 = 1.5, *x*32 = *x*42 = (2+3) / 2 = 2.5, *x*13 = *x*23 = (1+2) / 2 = 1.5, *x*14 = *x*24 = (2+3) / 2 = 2.5. The strong point of average method is easy to calculated mean value and its cost is low but it is based on the assumption that column is random variable but this thing is not totally exact because the null hypothesis requires that only *Xi* (s) being row vectors are random variables. On the contrary, mixture model approach follow hard on the requirement of null hypothesis.

Another way to predict empty values is to use Bayesian rule as estimation tool. It is better than average method because it takes advantage of probability distribution under 2*-*dimension data. Let *z* = *xuv* be the random variable that are dependent on row index *u* and column index *v* which, in turn, are random variables. According to Bayesian rule, the probability of *z* given *u* and *v* is:

Assuming that *u* and *v* are mutually independent given *z*, *P*(*z* | *u, v*) is re-written.

Let *θ* = *P*(*z*) = {*θ1,*  *θ2,…, θk*} be the parameter of *P*(*z* | *u, v*) where *θi* corresponds with the *ith* value of *z*, we have:

The goal of Bayesian estimate method is to determine *P*(*z* | *u, v*) via finding out optimal parameter *θ*. This method is iterative process whose iteration includes two steps: calculating *P*(*z* | *u, v*) based on *θ* and re-calculating *θ*. Let *t* and *θt* denotes *tth* iteration and the estimate of parameter *θ* at *tth* iteration. Note that *θ*0 can be initialized by 0.5. These two steps are described as below.

1. The current probability *P*(*z | x, y*) is calculated based on previous estimate of parameter *θt –*1.
2. Assigning current *P*(*z | x, y*) to current estimate of parameter *θt*.

If the deviation |*θt* – *θt –*1| is less than a threshold *ε*, then algorithm is stopped; otherwise go back step 1.

With example in table 2, suppose ranges of *u, v* and *z* are {1, 2, 3, 4}, {1, 2, 3, 4} and {1, 2, 3}, the conditional probabilities *P*(*u* | *z*) and *P*(*v* | *z*) are calculated and shown in following table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | *z* = 1 | *z* = 2 | *z* = 3 |
| *u* = 1 | 0.5 | 0.25 | 0 |
| *u* = 2 | 0 | 0.25 | 0.5 |
| *u* = 3 | 0.5 | 0.25 | 0 |
| *u* = 4 | 0 | 0.25 | 0.5 |
| *v* = 1 | 0.5 | 0.25 | 0 |
| *v* = 2 | 0 | 0.25 | 0.5 |
| *v* = 3 | 0.5 | 0.25 | 0 |
| *v* = 4 | 0 | 0.25 | 0.5 |

**Table III.3.** Conditional probabilities *P*(*u* | *z*) and *P*(*v* | *z*)

Suppose the parameter is initialized by 0.5, we have *θ*0 = *P*0(*z* | *u, v*) = 0.5. It means that the prior probabilities of *z* is *0.5*. It is easy to calculate the posterior probabilities *P*(*z* | *u, v*), for example:

By the similar way, the probabilities *P*(*z* | *u, v*), at the first iteration, over all values of *u, v* and *z* are calculated and shown in following table:

|  |  |  |  |
| --- | --- | --- | --- |
|  | *z* = 1 | *z* = 2 | *z* = 3 |
| *u*=1, *v*=1 | 0.8 | 0.2 | 0 |
| *u*=1, *v*=2 | 0 | 1 | 0 |
| *u*=1, *v*=3 | 0.8 | 0.2 | 0 |
| *u*=1, *v*=4 | 0 | 1 | 0 |
| *u*=2, *v*=1 | 0 | 1 | 0 |
| *u*=2, *v*=2 | 0 | 0.2 | 0.8 |
| *u*=2, *v*=3 | 0 | 1 | 0 |
| *u*=2, *v*=4 | 0 | 0.2 | 0.8 |
| *u*=3, *v*=1 | 0.8 | 0.2 | 0 |
| *u*=3, *v*=2 | 0 | 1 | 0 |
| *u*=3, *v*=3 | 0.8 | 0.2 | 0 |
| *u*=3, *v*=4 | 0 | 1 | 0 |
| *u*=4, *v*=1 | 0 | 1 | 0 |
| *u*=4, *v*=2 | 0 | 0.2 | 0.8 |
| *u*=4, *v*=3 | 0 | 1 | 0 |
| *u*=4, *v*=4 | 0 | 1 | 0 |

**Table III.4.** Posterior probabilities of *z*, *P*(*z | u, v*)

At next iteration posterior probabilities *z*, *P*(*z | u, v*) are computed by similar way except that the parameter *θ* is assigned by values in table 3. In other words, the parameter for next iteration is assigned as current posterior probabilities *z*, *P*(*z | u, v*). Now we estimate missing values such as *x*31, *x*41, *x*32, *x*42, *x*13, *x*23, *x*14 and *x*24. The posterior probabilities for *x*31 = 1, *x*31 = 2, *x*31 = 4 are *P*(*z* = 1 | *u* = 3, *v* =1) = 0.8, *P*(*z* = 2 | *u* = 3, *v* =1) = 0.2 and *P*(*z* = 3 | *u* = 3, *v* =1) = 0, respectively. Because *P*(*z* = 1 | *u* = 3, *v* =1) = 0.8 is maximal, *x*31 receives value 1, *x*31 = 1. By the similar way, we have *x*41 = 2, *x*32 = 2, *x*42 = 3, *x*13 = 1, *x*23 = 2, *x*14 = 2 and *x*24 = 3. These estimates are much better than ones resulted from average method even if the algorithm have just come over one iteration. Following is the complete sample whose missing values are estimated.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *v* = 1 | *v* = 2 | *v* = 3 | *v* = 4 |
| *u* = 1 | *x*11 = 1 | *x*12 = 2 | *x*13 = 1 | *x*14 = 2 |
| *u* = 2 | *x*21 = 2 | *x*22 = 3 | *x*23 = 2 | *x*24 = 3 |
| *u* = 3 | *x*31 = 1 | *x*32 = 2 | *x*33 = 1 | *x*34 = 2 |
| *u* = 4 | *x*41 = 2 | *x*42 = 3 | *x*43 = 2 | *x*44 = 3 |

**Table III.5.** Complete sample whose missing values are estimated by Bayesian method

However, the drawback of Bayesian estimate method in hypothesis testing is that it gives three assumptions as below.

* Sample scalar values *xij* (s) are considered as random values while vector *Xi* (s) are random variables in null hypothesis. Scalar distribution is different from vector distribution with regard to the modality of data. When the modality of data is transformed, the meaning of data is changed unpredictably.
* Row index *u* and column index *v* are also considered as random variables but this assumption doesn’t exist in null hypothesis.
* Row variable *u* and column variable *v* are mutually independent given sample value *xij*.

In general, mixture model approach is better than filling missing values approaches in flavor of hypothesis testing because it does not give any additional assumption about sample space except assumptions that null hypothesis issues such as the normality of sample data. In other words, it conserves attributes of sample data.

## Conclusion

In general, the basic idea of this research is to analyze the global distribution as a set of partial (or potential) distributions. Because it is impossible to apply the global distribution into incomplete sample, the research applies each partial distribution into a sub-sample of incomplete sample where sub-sample is defined as a piece of incomplete sample so that such sub-sample contains complete observation vectors. Partial distributions are unified into a mixture model and expectation maximization (EM) algorithm is used to find out optimal parameters of mixture model such that these optimal parameters maximize the likelihood function. The ratio of the maximum likelihood of null hypothesis to the maximum likelihood of alternative hypothesis is totally determined based on these optimal parameters. Consequently, this ratio is used to test null hypothesis in flavor of alternative hypothesis by the normal way that is to take advantage of chi-square distribution. The excellent of this research is that there is no requirement of filling missing values existing in incomplete sample. When missing values are replaced by estimated values, inherent attributes of sample data changed or disturbed even if such estimated values are considered as the best-predicted values. In other words, this method maintains inherent attributes of sample data. Moreover, the higher the density of missing values is, the more effective this research is. Especially, if the data incompletion occurs frequently or conforms a particular period or rule in process of collecting samples, then the corporation of partial distributions and likelihood ratio test delivers the best result. However if missing values occur randomly such as pepper-and-salt noise, this approach is not good choice for test because pepper-and-salt noise data has no inherent features and so partial distributions cannot be specified and estimated precisely.

# Multivariate non-parametric hypothesis testing

Non-parametric testing is very necessary in case that the statistical sample does not conform normal distribution or we have no knowledge about sample distribution. Sign test is a popular and effective test for non-parametric model but it cannot be applied into multivariate data in which observations are vectors because the ordering and comparative operators are not defined in *n-*dimension vector space. So, this research proposes a new approach to perform sign test on multivariate sample by using a hyper-plane to separate multi-dimensional observations into two sides. Therefore, it is possible for the sign test to assign plus signs and minus signs to observations in each side. Moreover, this research introduces a new method to determine the separated hyper-plane. This method is a variant of support vector machine (SVM), thus, the optimized hyper-plane is the one that contains null hypothesis and splits observations as discriminatively as possible.

## A proposal of non-parametric sign test in multivariate data

As aforementioned in previous section, nonparametric testing is used in case of without knowledge about sample distribution; concretely, there is no assumption of normality. The nonparametric testing begins with the test on sample *median* in univariate data. If distribution is symmetric, median is identical to mean. Given the median is the observation at which the left side data and the right side data are of equal accumulate probability.

*P*(*D <* ) *= P*(*D >* ) *=* 0.5

If data is not large and there is no assumption about normality, the median is approximate to population mean. Given null hypothesis *H*0: = and alternative hypothesis *H*1: , the so-called sign test is performed as below steps (Walpole, Myers, Myers, & Ye, 2012, pp. 656-660):

* Assigning plus signs to sample observations whose values are greater than and minus signs to ones whose values are less than . Note that values which equal are not considered. Plus signs and minus signs represent the right side and left side of , respectively.
* If the number of plus signs is nearly equal to the number of minus signs, then null hypothesis *H*0is true; otherwise *H*0 is false. In other words, that the proportion of plus signs is significantly different from 0.5 cause to rejecting *H*0 in flavor of *H*1.

The reason of *H*0 acceptance is that the probability that observations fall in both left side and right side of are of equal value 0.5 and of course, it is asserted that is a real median. Note that terms *data point*, *sample point*, *sample value* and *observation* are identical.

If sample is extended from real number space to *p-*dimension space, observations become vectors composed of *p* partial values. Traditional sign test cannot be applied into multivariate data because null hypothesis and alternative hypothesis are vectors and it is not likely to compare hypothesis vector with observation vector with regard to assign plus signs. This paper proposes a new approach for sign test in case of *p-*dimension sample. The basic idea is to determine *a hyper-plane and its normal vector which is used to calculate signed-rank number*, namely, an observation is assigned plus sign if it is in the right side of hyper-plane. The right side, in turn, is defined by the direction of normal vector of hyper-plane. Because the hyper-plane is very important, research also proposes a method to determine it with proposition “the optimized hyper-plane is the one that contains null hypothesis and splits observations as discriminatively as possible”.

Suppose distribution is symmetric, median is identical to mean, we have null hypothesis *H*0: *μ* = *μ*0 and alternative hypothesis *H*1: *μ*  *μ*0, the sign test in case of multivariate sample includes three following steps:

1. Determining separated hyper-plane that contains null hypothesis *μ*0 and splits observations as discriminatively as possible. The method to discover this hyper-plane is described particularly in next section. Another important thing is to specify the normal vector *W* of such hyper-plane. The normal vector *W* is used to make sign of observations.
2. Let Ω be *p-*dimension multivariate sample and let *Xi* = (*xi*1, *xi*2,…, *xip*) Ω be observation vector. The scalar product *WT*(*Xi – μ*0) between normal vector *W* and each standardized observation *Xi – μ*0 is calculated, where *T* denotes transposition operator and *W* and *Xi* are column vectors. If such scalar product is positive then *Xi* is assigned a plus sign; otherwise *Xi* is assigned a minus sign.
3. Let *b* is the number of plus signs, thus, *b* conforms binomial distribution with proportion parameter 0.5. Let *z* be the *z*-score of *b*, we have *z* = where *n* is the number of observations in sample, *n =* |Ω|. If *|z| > zα*/2then rejecting *H*0 in flavor of *H*1 where *zα*/2 is 100(1 *– α*/2) percentage point of standard normal distribution. It is easy to recognize that *z* is normal approximation of *b*.

The ideology behind this method is that if vector *μ*0 is the mean of symmetric multivariate distribution, then it exists a separated hyper-plane that divides the whole sample space into two half-spaces so that the cardinalities of these half-spaces are equal (or approximately equal). It is possible to exist many hyper-planes like that but the optimal hyper-plane is found out by the method described particularly in next section.

## A new method to determine separated hyper-plane

Let Ω be *p-*dimension multivariate sample of size *n* and let *Xi* = (*xi1, xi2,…, xip*) Ω be observation vector. Suppose Ω = {*X1, X2,…, Xn*} has *n* observations. The method is to find out an optimal separated hyper-plane is a variant of support vector machine (SVM) (Law, 2006). Hence, this hyper-plane satisfies two following propositions:

1. Containing specified null hypothesis *μ*0.
2. *S*plitting vector space Ω into two half-spaces as separately as possible.

The first proposition is by following equation that is the equation of hyper-plane.

*WT*(*Xi – μ*0) *= 0*

Where *W* is the normal vector of hyper-plane and *T* denotes transposition operator. The task of determining optimal hyper-plane is identical to find out its normal vector *W.* The second proposition is equivalent to maximizing the distance between two margins of two half-spaces. This proposition is similar to the methodology of support vector machine (SVM) method. These margins between two half-spaces are restricted by two parallel hyper-planes that are specified by following equations:

Two half-spaces are separated as much as possible if and only if following condition is satisfied:

Because the distance between two margins is , the optimal hyper-plane will maximize . It means that the optimal hyper-plane minimizes with constraint . According to Lagrange dual theorem (Boyd & Vandenberghe, 2009, p. 215), *W* is extreme point of following Lagrange function:

Where *λ =* (*λ*1, *λ*2,…, *λn*)*T* represents a set of Lagrange multipliers and . Suppose *W* minimizes then it is the solution of following equation when we set the derivative of *L*(*W*, *α*) with regard to *W* to 0.

Substituting (IV.2) into equation (IV.1), we have:

In case of ; due to and , we have:

In case of ; due to and , we have:

Due to

In any case we have:

It is easy to infer that

So *L*1(*λ*) and *L*2(*λ*) are inferior and supreme of *L*(*W, λ*), respectively.

*L*1(*λ*) *≤ L*(*W, λ*) *≤ L*2(*λ*)

Hence, *L*1(*λ*) and *L*2(*λ*) which are functions of Lagrange multipliers are re-written into matrix notation, according to equation (IV.3):

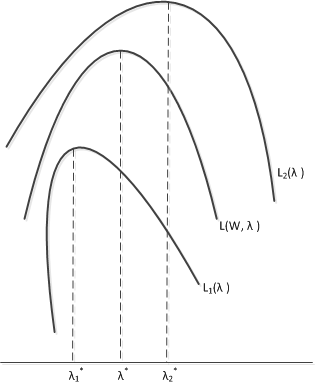
(**IV.3**)

Where *I* = (1*,* 1*,…,* 1)*T* is one-column identity matrix and *S* is a symmetric *n*x*n* matrix with elements *sij =* . Minimizing is identical to maximizing *L*(*W*, *λ*) with regard to *λ*. Note that maximizing *L*(*W*, *λ*) is equivalent to maximizing *L*1(*λ*) and *L*2(*λ*) because *L*1(*λ*) and *L*2(*λ*) are inferior and supreme of *L*(*W, λ*), respectively. Suppose *λ*1*\** and *λ*2*\** is maximum points of *L*1(*λ*) and *L*2(*λ*), respectively, we have equation (IV.4) for specifying them:

(**IV.4**)

*L*1(*λ*) and *L*2(*λ*) are second-order functions and so the way to find *λ*1*\** and *λ*2*\** is quadratic programming (QP) problem. There are many approaches to solve this problem; for instance, the sequential minimal optimization is described in (Law, 2006, p. 15) and (Platt, 1998, pp. 8-9).

Although *λ*1*\** and *λ*2*\** is maximum points of *L*1(*λ*) and *L*2(*λ*), they are not asserted to be maximum points of *L*(*W*, *λ*) with regard to *λ*. Let *λ\** be the maximum point of *L*(*W*, *λ*) with regard to *λ*, following figure is the interpretation of *λ*1\*, *λ*2\*, *λ*\*, *L*1(*λ*), *L*2(*λ*) and *L*(*W*, *λ*).



**Figure IV.2.1.** The interpretation of *λ*1*\**, *λ*2*\**, *λ\**, *L*1(*λ*), *L*2(*λ*) and *L*(*W*, *λ*)

So the maximum point *λ\** is approximated by the average of *λ*1*\** and *λ*2*\** according to equation (5).

Substituting maximum point *λ \** to equation (IV.2), we have following equation:

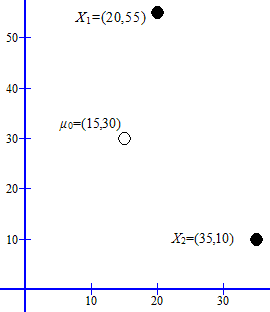
Because equation (IV.6) is the quadratic function, it is easy to find out its solutions. There are some methods to solve quadratic equation such as Newton-Raphson and bisection (Burden & Faires, 2011, pp. 48-74). If equation (IV.6) has only one solution *W\**, then *W\** is normal vector of separated hyper-plane. Otherwise equation (IV.6) has *k* > 1 solutions *W*1*\**, *W*2*\**,…, *Wk\**, then the normal vector *W\** of separated hyper-plane is the addition of these solutions.

So the best normal vector *W\** is considered as the intermediate one among possible normal vector (*W*1*\**, *W*2*\**,…, *Wk\**). As aforementioned, it is easy to determine the separated hyperplane with regard to *W\** according to equation (IV.7).

Where *μ*0 is the null mean (median).

## A case study of non-parametric sign test based on separated hyperplane

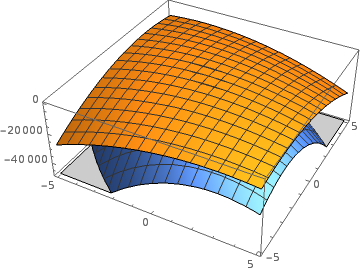
Given a sample including two data points *X*1={20, 55}, *X*2={35, 10} and a null median , it is necessary to perform a sign test with null hypothesis *H*0: = in flavor of alternative hypothesis *H*1: ≠ . Figure IV.3.1 shows points *X*1, *X*2 and median .



**Figure IV.3.1.** Data sample and null median

According to equation (3), inferior and supreme of *L*(*W*, *λ*) are:

Figure IV.3.2 shows 3-dimension graphs of *L*1(*λ*) and *L*2(*λ*).



**Figure IV.3.2.** Graphs of *L*1(*λ*) and *L*2(*λ*)

The maximum points of *L*1(*λ*) and *L*2(*λ*) are found by setting their partial derivatives to be zero. We have:

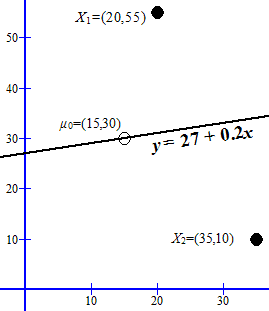
So *L*1(*λ*) gets maximal at . Similarly, we have:

So *L*2(*λ*) gets maximal at . According to equation (5), the maximum point *λ\** is the average as follows:

By substituting *λ\** into equation (6) where *W*=(*w*1, *w*2), we have (Wolfram):

The normal vector is solution of the above equation. The separated hyperplane is specified based on *W*\* according to equation (7) as follows:

Figure IV.3.3 shows the separated hyperplane *y* = 27 + 0.2*x* with regard to two variables.



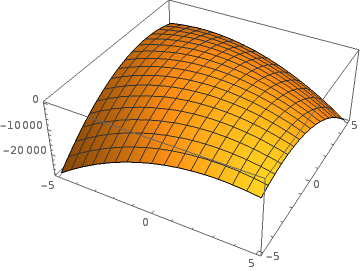
**Figure IV.3.3.** Separated hyperplane with regard to two variables

Due to

There are 1 positive sign and 1 negative sign. According to sign test, that the number of plus signs is equal to the number of minus signs leads to fail to reject the null hypothesis *H*0: *μ*=(15,30).

Moreover, if we substitute into equation (1), the Lagrange function *L*(*λ*) is the same to supreme *L*2(*λ*) as follows:

Figure IV.3.4 shows 3-dimension graph of *L*(*λ*).



**Figure IV.3.4.** Graphs of *L*(*λ*)

## Conclusion

The main idea of this paper is to find out a separated hyper-plane, which aims to count the number of observations which fall in left side or right side of such hyper-plane. The right side of hyper-plane is defined as the side that its direction is the same to normal vector of hyper-plane. The opposite to the right side is the left side. The method to find out hyper-plane is a variant of support vector machine (SVM) [Cristianini, Shawe-Taylor 2000] method except that hyper-plane contains the null hypothesis. SVM assumes that data points are labeled with classes {1 and *–*1} but observations in case of hypothesis testing are not classed. This research solves this problem by two-step process of separation:

* Firstly, the label of observations is specified indirectly via the scalar product between these observations and the normal vector of separated hyper-plane.
* Secondly, Lagrange multipliers are determined via arithmetic transformations when the goal of such transformations is to eliminate normal-vector-dependency from Lagrange function and find out the inferior and supreme of Lagrange function. After that the best-separated hyper-plane is the intermediate one among many possible hyper-planes.

# Conclusion and future trends